NERSC Update

Horst D. Simon
NERSC
Lawrence Berkeley National Laboratory

ASCAC Meeting, Washington D.C.
August 9, 2006

BERKELEY LAB

NERSC

Office of Science
U.S. DEPARTMENT OF ENERGY

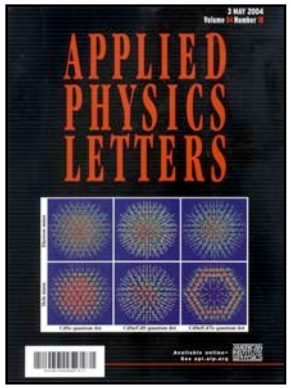
Overview

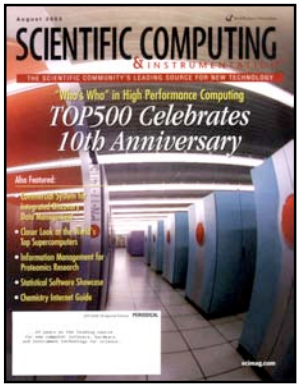
- New NERSC systems in 2005 and 2006
- Impact on science
- Plans for 2006 and 2007: NERSC 5

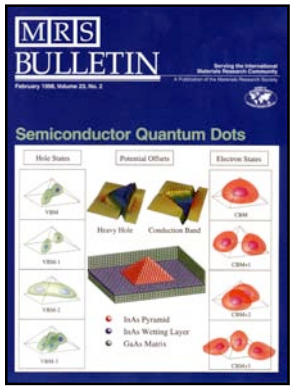
NERSC Mission

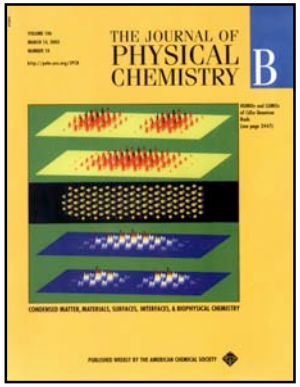
The mission of the National Energy Research Scientific Computing Center (NERSC) is to accelerate the pace of scientific discovery by providing high performance computing, information, data, and communications services for research sponsored by the DOE Office of Science (SC).

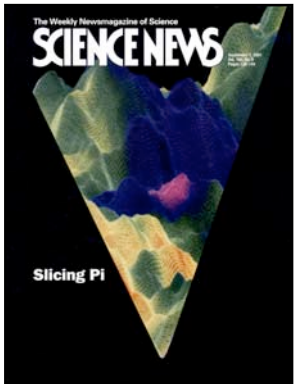
Science-Driven Computing

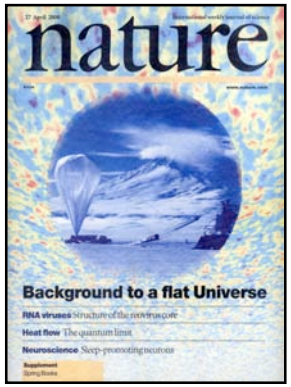


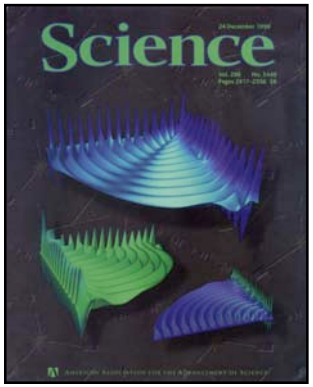


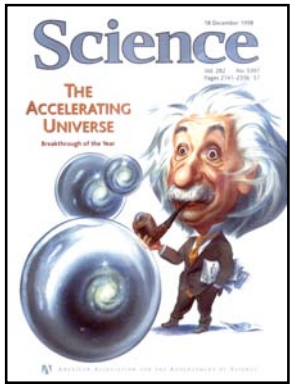












NERSC is enabling new science

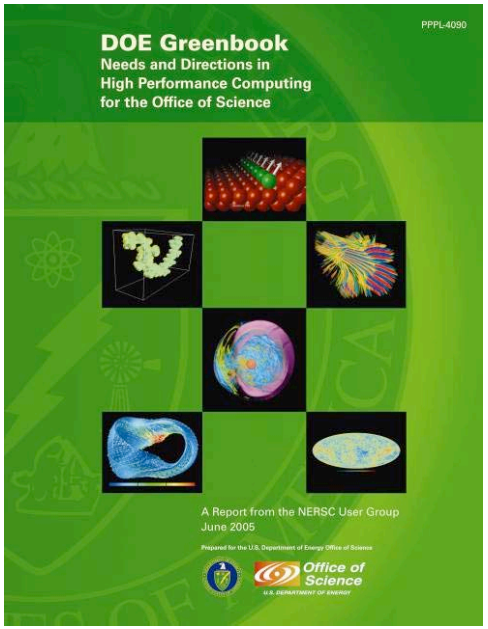
NERSC

BERKELEY LAB

Office of Science
U.S. DEPARTMENT OF ENERGY

2005: NERSC Greenbook

- **Edited by S. Jardin (PPPL) with contributions by 37 scientists**
- **Recommendations for the next five years:**
  - **expand HPC resources, maintaining a system balance to support the wide range of applications in SC**
  - **configure the systems to minimize the time-to-completion of large jobs; maximize the overall efficiency of the hardware**
  - **actively support the continued development of algorithms, software, and database technology for improved performance on parallel platforms**
  - **strengthen the computational science infrastructure at NERSC that will enable the optimal use of current and future NERSC supercomputers**
  - **evaluate the requirements of data- or I/O-intensive scientific applications in order to support as wide a range of science as possible.**



NERSC BERKELEY LAB Office of Science U.S. DEPARTMENT OF ENERGY

2005: NERSC Five Year Plan

- **Three trends that need to be addressed:**
  - **the widening gap between application performance and peak performance of high-end computing systems**
  - **the recent emergence of large, multidisciplinary computational science teams in the DOE research community**
  - **the flood of scientific data from both simulations and experiments**
- **Requirements and Trend Analysis then led to the development of the NERSC Strategic Plan**

Science-Driven Computing Strategy 2006 -2010

DOE Review of NERSC, May 2005

- **Programmatic Review chaired by F. Williams (Arctic SC Center) with panel of eight reviewers**
- **Specific questions, but could look into all areas of operations**
- **Strong endorsement of NERSC's approach to fullfilling its mission**
  - "NERSC is a strong, productive, and responsive science-driven center that possesses the potential to significantly and positively impact scientific progress …."
  - "… NERSC is extremely well run with a lean and knowledgeable staff. The panel members saw evidence of strong and committed leadership, and staff who are capable and responsive to users' needs and requirements. Widespread, high regard for the center's performance, reflected in such metrics as the high number of publications supported by NERSC, and its potential to positively impact future advancement of computational science, warrants continued support."

New Production Resource (08/2005): "Jacquard"

- 722-processor (AMD 2.2 GHz Opteron) Linux Networx Evolocity cluster
- one of the largest production InfiniBand-based systems
- met rigorous acceptance criteria for performance, reliability, and functionality unprecedented for an InfiniBand cluster
- first system to deploy Mellanox 12x InfiniBand uplinks in fat-tree

New Production Resource (01/2006): "Bassi"

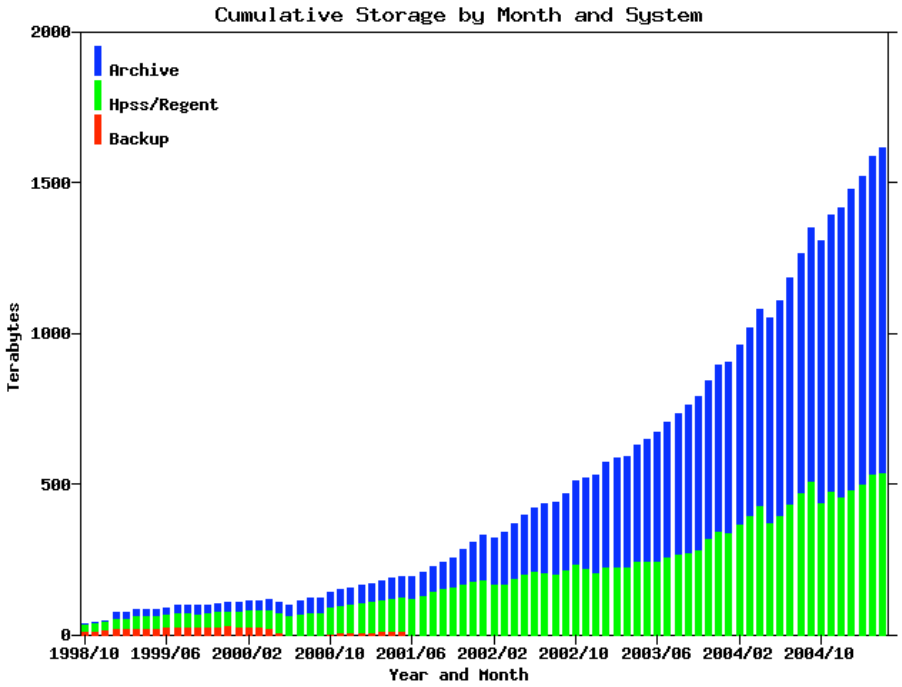
- **122 IBM p5-575 nodes (with 32GB each)**
- **1.9 GHz POWER 5 processors**
- **111 compute nodes (888 processors)**
- **Dual plane Federation interconnect**
- **7.6 GFlops/sec peak proc. speed**
- **100 TB of usable disk space in GPFS**
- **accepted and installed in record time**

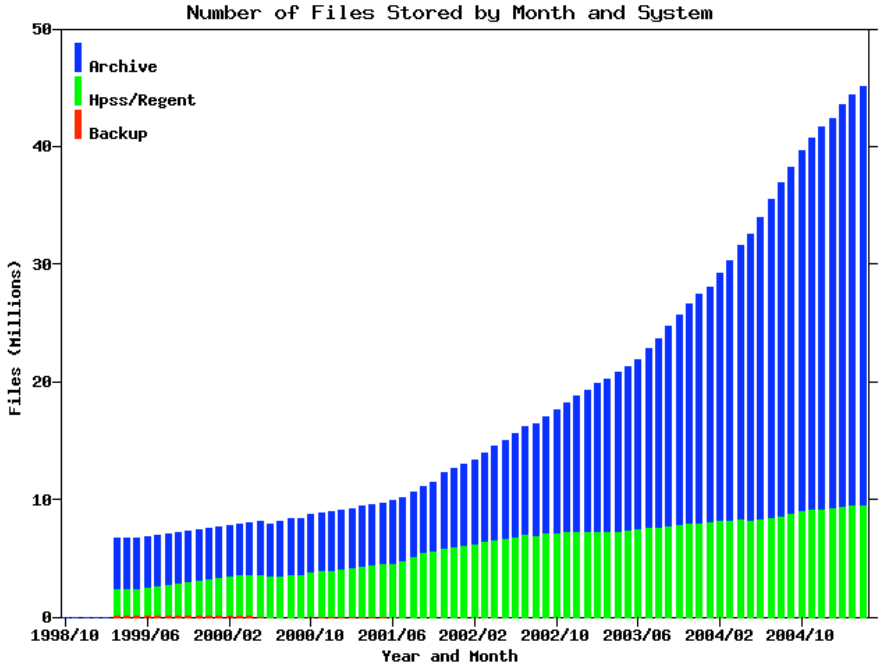
"I have to say that both of these machines are really nothing short of fabulous … While Jacquard is perhaps the best-performing commodity cluster I have seen, Bassi is the best machine I have seen, period."
--- Robert Duke, UNC Chapel Hill

2006: NERSC Global Filesystem (NGF)

- After thorough evaluation and testing phase in production since early 2006
- Based on IBM GPFS
- Seamless data access from all of NERSC's computational and analysis resources
- Single unified namespace makes it easier for users to manage their data across multiple system
- First production global filesystem spanning five platforms, three architectures, and four different vendors

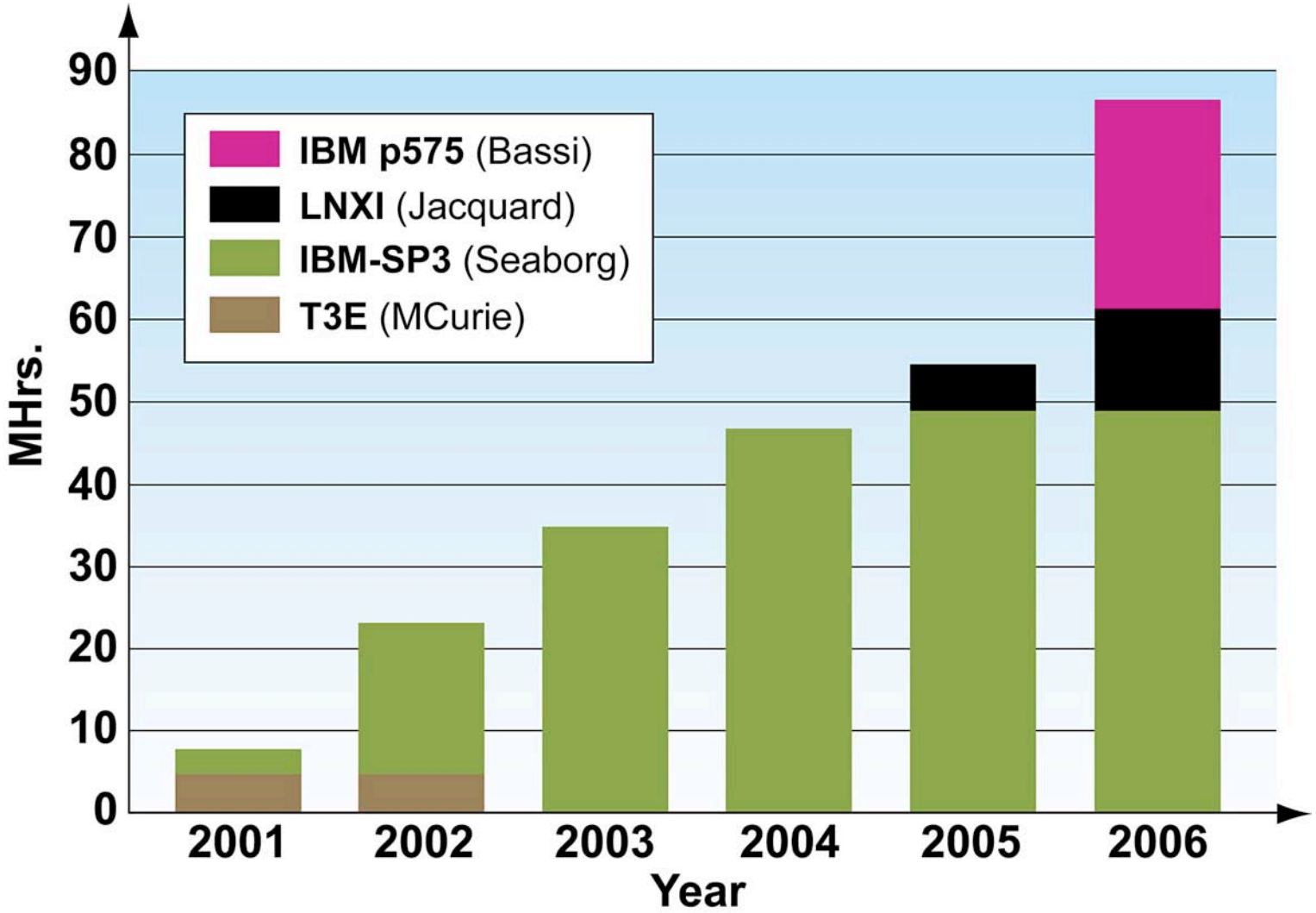
NERSC: one of the largest open science storage environments





45 million files
44 PB capacity
1.7x per year data growth

NERSC Allocations

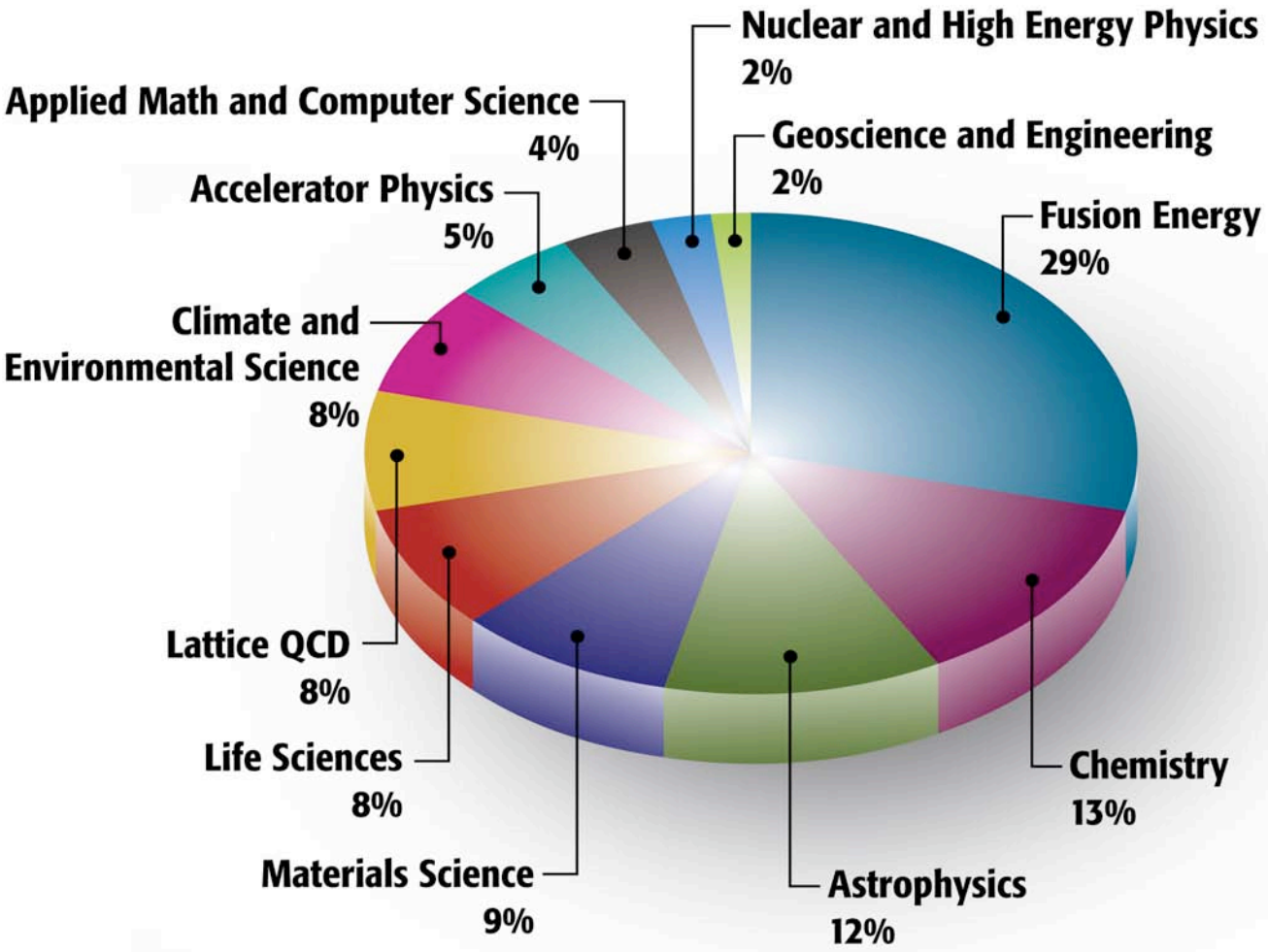


NERSC

Berkeley Lab

Office of Science
U.S. Department of Energy

Usage by Discipline (2005)



Impact on Science Mission

Acknowledgments

A.A.G. wishes to thank Roland Assaraf for validating Zori against QMCMOL. A.A.G. also thanks Anthony Scemama for his contribution of electron pair localization function routines. Computer time was provided by the Department of Energy's Innovative and Novel Computational Impact on Theory and Experiment (INCITE) program. D.D. was supported by the CREST Program of the National Science Foundation under Grant No. HRD-0318519.

P. N. and D. K. acknowledge support from a NASA LTSA and ATP grant. This research used resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the US Department of Energy under contract DE-AC03-76SF00098. We thank them for a generous allocation of computing time under the "Big Splash" award, without which this research would have been impossible.

We thank the RHIC Operations Group and RCF at BNL, and the NERSC Center at LBNL for their support. This work was supported in part by the HENP Divisions of the Office of Science of the U.S. DOE; the

Acknowledgements

H.W. thanks the National Energy Research Scientific Computing Center (NERSC), which is supported by the Office of Science of the US Department of Energy under Contract No. DE-AC03-76SF00098, for the allocation of computer time. M.T. thanks W. Domcke, M. Gelin, A. Pisliakov, and G. Stock for numerous helpful discussions. This work has been supported in part by a

The authors are grateful to Prof. R. J. Bartlett and Dr. M. Musial, who very kindly computed the CISDTQ and CCSDTQ numbers reported in this work. G. K-L. Chan would also like to thank Prof. N. C. Handy, who, as always, pointed him in the right direction. Most of the computations were carried out at the NERSC supercomputer centre, via DOE grant 12345, and the NERSC staff (in particular D. Skinner) are thanked for their assistance in many technical matters.

Acknowledgements This work was supported by the US Department of Energy and the National Science Foundation and used resources of the National Energy Research Scientific Computing Center at LBNL; C.G. was also supported by the Hertz Foundation. C.G. acknowledges his faculty advisor J. Wurtele. We appreciate contributions from G. Dugan, J. Faure, G. Fubiani, B. Nagler, K. Nakamura, N. Saleh, B. Shadwick, L. Archambault, M. Dickinson, S. Dimaggio, D. Syversrud, J. Wallig and N. Ybarrolaza.

Impact on Science Mission

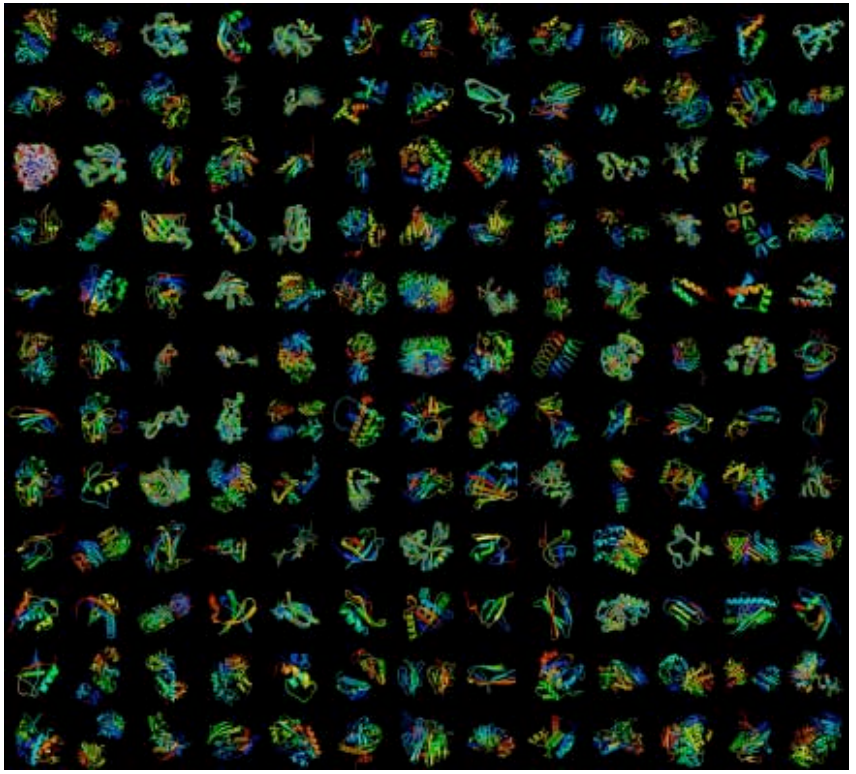
- Majority of great science in SC is done with medium- to large-scale resources
- In 2005, NERSC users reported the publication of more than 1200 papers that were based wholly or partly on work done at NERSC.

NERSC Supports

- Different types of projects
  - Single PI projects
  - Large computational science collaborations
  - INCITE
- Large variety of applications
  - All scientific applications in DOE SC
- Range of Systems
  - Computational, storage, networking, analytics

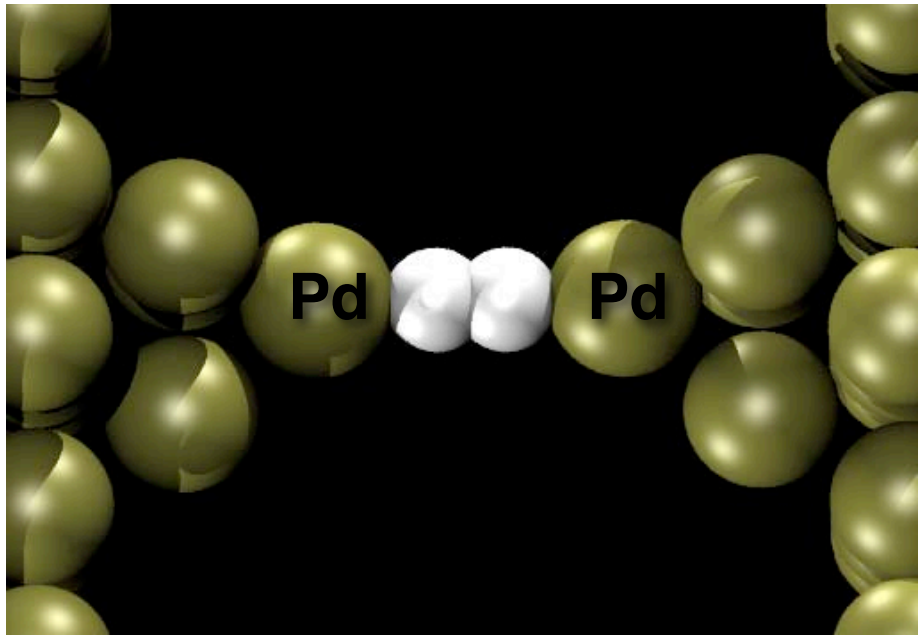
INCITE Project: Molecular Dynameomics

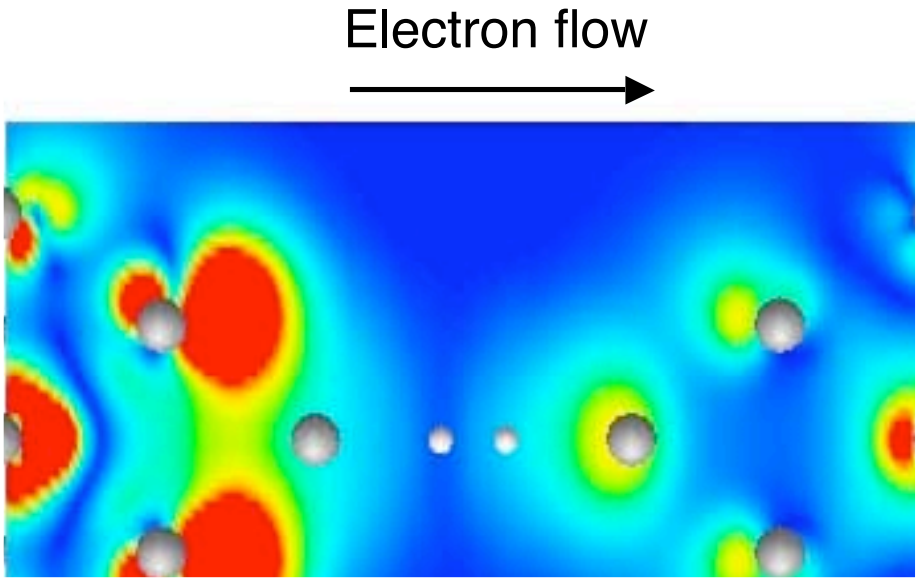
- V. Daggett, U Washington, 2M NERSC proc. hours
- Understand protein folding pathways by 'unfolding' proteins at high temp.
- Computed unfolding of 151 most common fold structures at different temperatures
- Multiple runs of MD calculation for each fold/temp. pair



The first 156 protein targets

Exploring the Limits of Nanoelectronics with Theory: Single Molecule Electrical Junctions

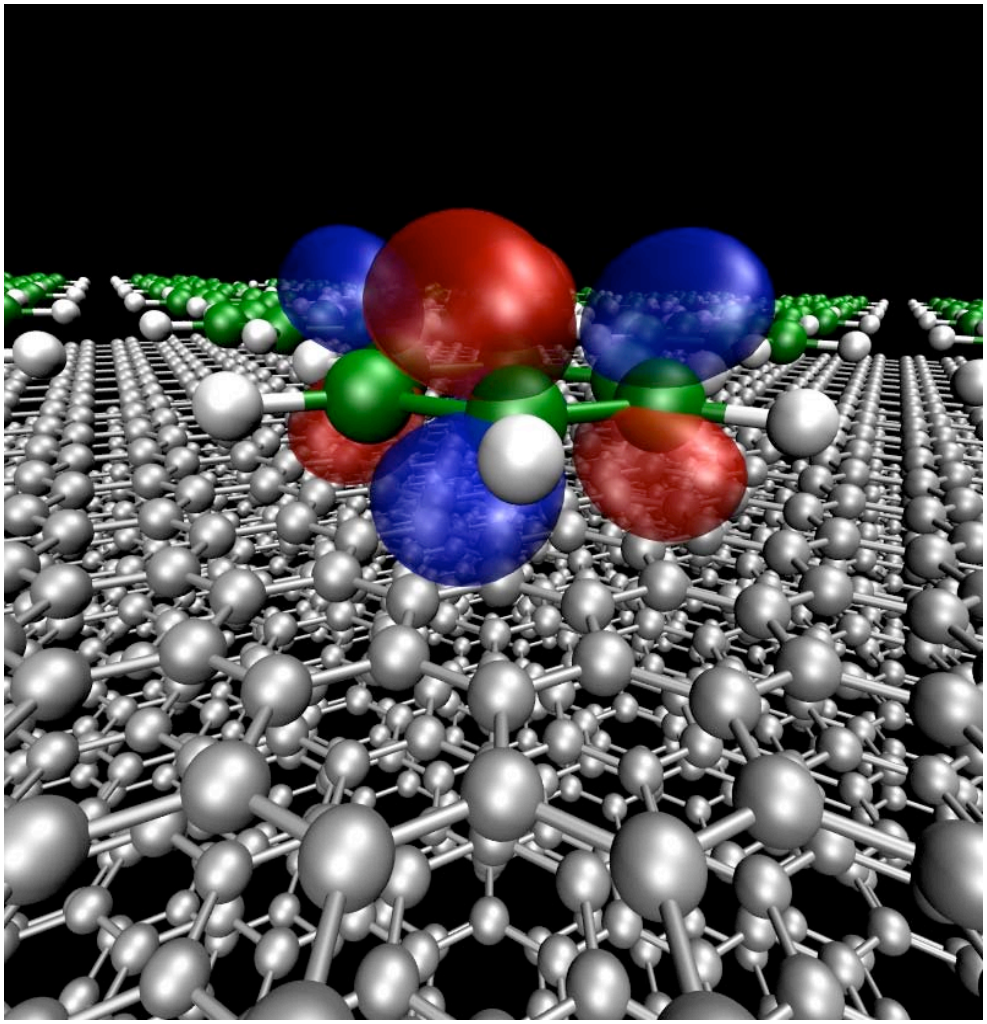




Shown at left is a single hydrogen molecule (in white) bridging palladium point contacts. At right, a density plot of the dominant transmitting electronic state reveals a significant reflection of charge at the left Pd contact, leading to a high resistance, consistent with recent experiments. (Red is high electronic density in the plot, blue is low.)

Steven Louie,
Marvin Cohen,
UC Berkeley
Jeff Neaton,
Molecular Foundry

Excited electronic states at metal-organic interfaces

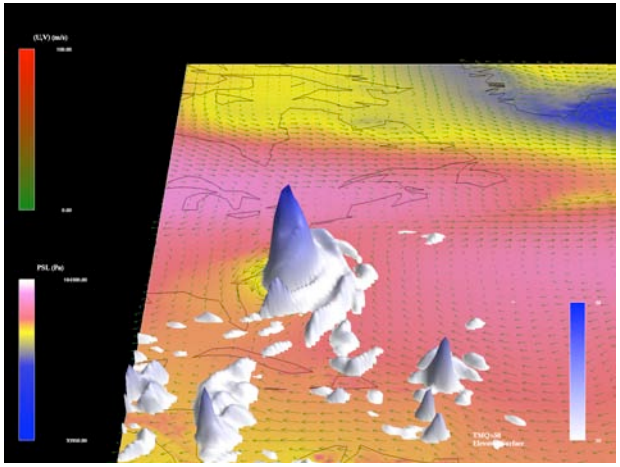


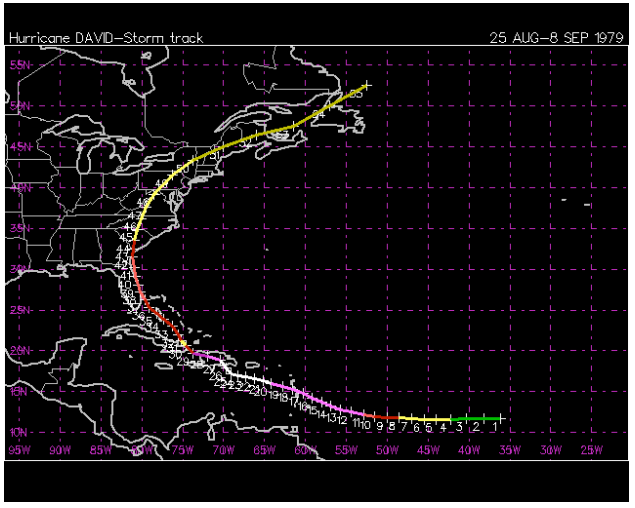
Mark Hybertsen &
George Flynn
Columbia University
Jeff Neaton
Molecular Foundry

Lowest unoccupied molecular orbital of a benzene molecule physisorbed on a graphite surface. Our calculations predict that, relative to the gas-phase, orbital energies are strongly modified by the surface.

NERSC Science Driven Analytics: Comparing Real and Simulated Storm Data

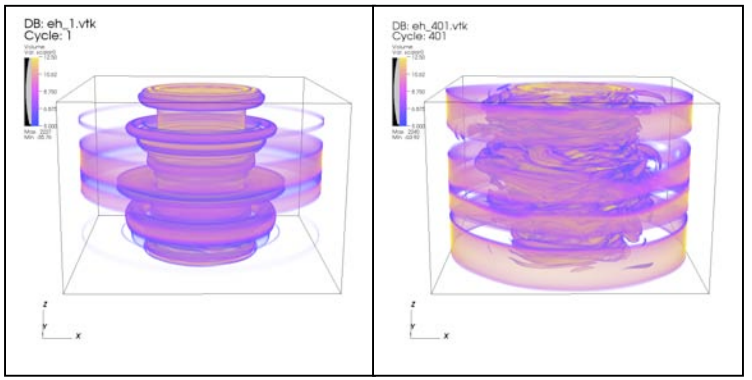
- NERSC designed a prototype workflow enabling fast qualitative comparisons between simulated storm data and real observations
- By using the NERSC Global Filesystem (NGF) the most appropriate resource can be used at each stage
  - IBM P5 (Bassi) for large-scale parallel computing
  - Linux cluster (Jacquard) for data reduction
  - Visualization server

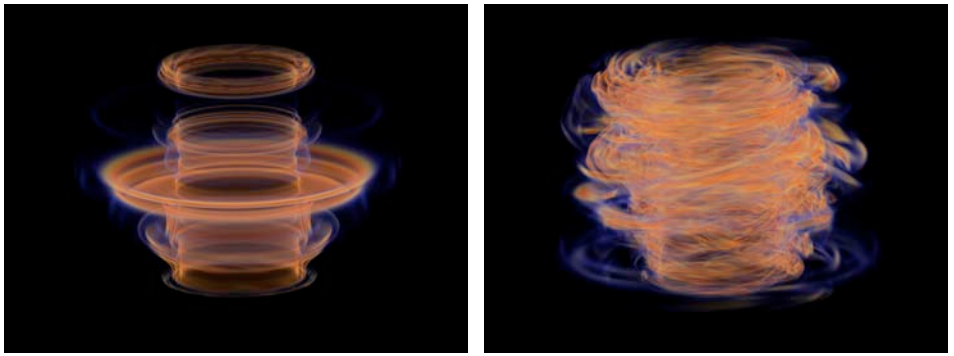


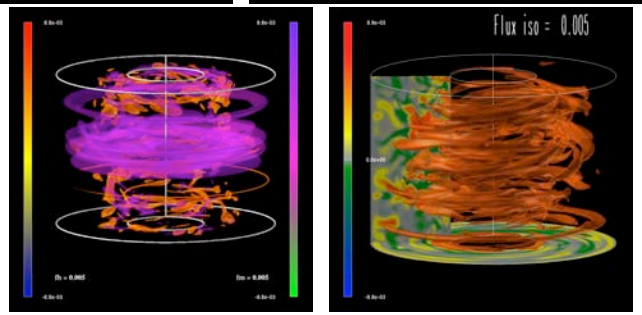


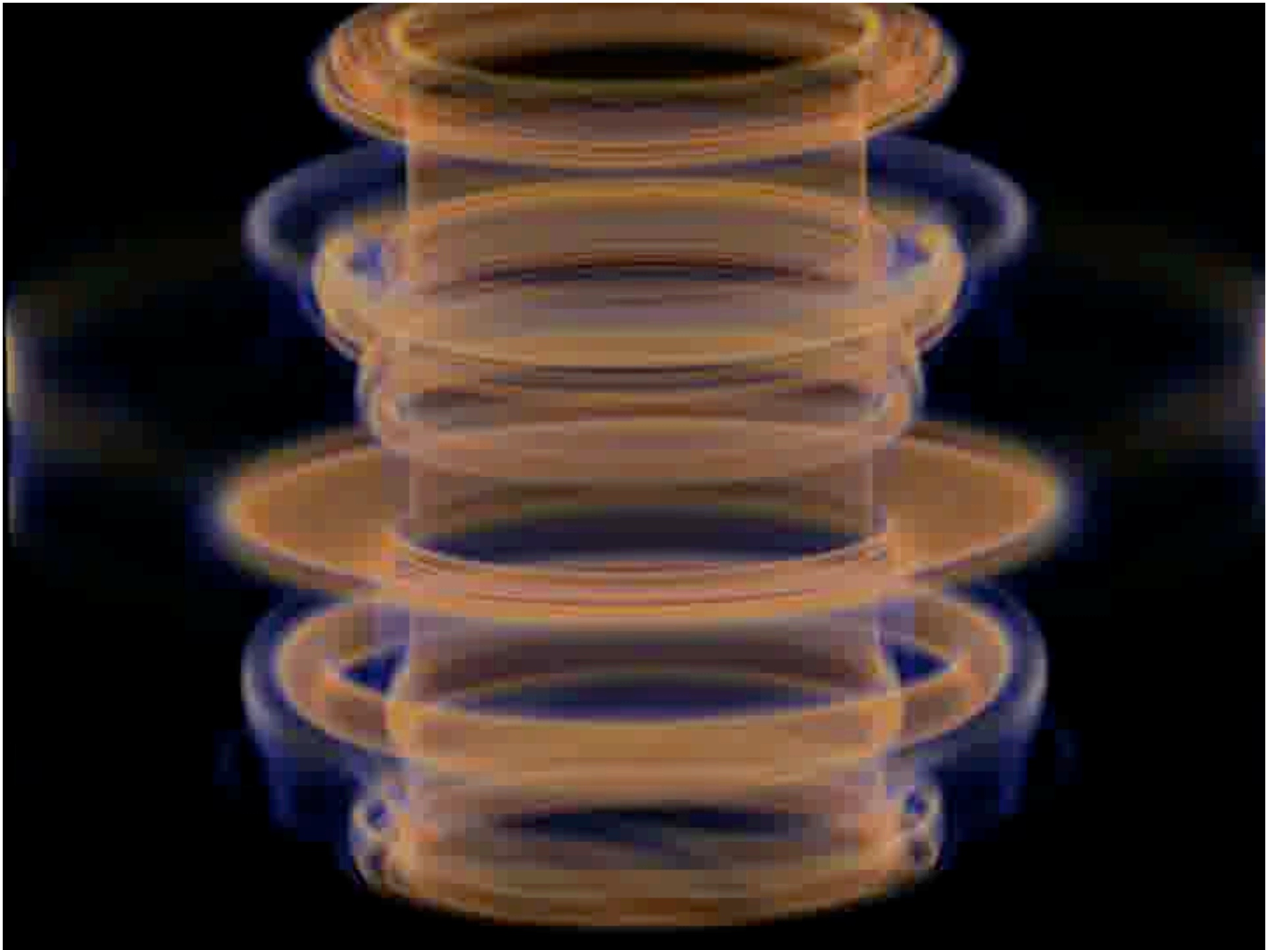
INCITE 4 – Magneto-Rotational Instability and Turbulent Angular Momentum Transport

- Visual Analytics support for collaboration with F. Catteano, Univ. of Chicago
- iterative, investigatory approach to explore alternative methods in order to determine which one provides the best visual and scientific results.
- Top row: hydro enstrophy from two different timesteps.
- Middle row: magnetic enstrophy from two different timesteps
- Bottom row: hydro and magnetic flux (left), total advective radial flux of axial angular momentum (right).
- Movie of time-evolving magnetic enstrophy.



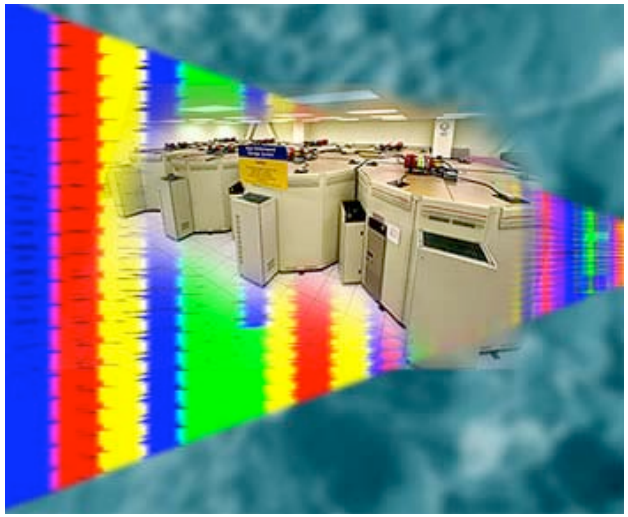






Archiving for Genomics Research

- **Production Genome Facility (PGF) at Joint Genome Institute (JGI) is producing sequence data at increasing rate**
  - **2 million files per month of trace data (25 to 100 KB each)**
  - **100 assembled projects per month (50 MB to 250 MB)**
  - **several very large assembled projects per year (~50 GB).**
  - **total about 2 TB per month on average**
- **NERSC and PGF staff collaborated to set up data pipeline for nightly back-up, using ESnet's new Bay Area MAN**



NERSC 5

NERSC-5: Nationally Coordinated Procurement Process

- The HECRTF and NRC Reports recommend coordination of Government procurements
- NERSC-5 is possibly the first procurement to coordinate with other agencies
  - Joint application and kernel benchmarks with DOD HPCMP TI-06
  - Joint application and kernel benchmarks with NSF
  - NERSC-5 evaluation had four organizations observing the process
- NERSC had considerable influence on NSF Petascale procurement process

NERSC-5 Goals

- **Sustained System Performance over 3 years**
  - **7.5 to 10 sustained Teraflop/s averaged over 3 years**
- **System Balance**
  - **Aggregate memory**
    - **Users have to be able to use at least 80% of the available memory for user code and data.**
  - **Global usable disk storage**
    - **At least 300 TB with an option for 150 TB more a year later**
  - **Integrate with the NERSC Global Filesystem (NGF)**
- **Expected to significantly increase computational time for NERSC users in the 2007 Allocation Year**
  - **Dec 1, 2006 – November 30, 2007**
  - **Have full impact for AY 2008**
  - **Can arrive in FY 2006**

Greenbook and Other Plans

- Coordinated requirements with the NUG Greenbook
- Aligned with the NERSC five year plan for 2006-2010

NERSC-5 Benchmarks

- **Selection of benchmarks - several considerations**
  - **Representative of the workload**
  - **Represent different algorithms and methods**
  - **Are portable to likely candidate architectures with limited effort**
  - **Work in a repeatable and testable manner**
  - **Are tractable for a non-expert to understand**
  - **Can be instrumented**
  - **Authors agree we can use and distribute it**
- **NERSC-5 started with approximately 20 candidates – settled on 7**

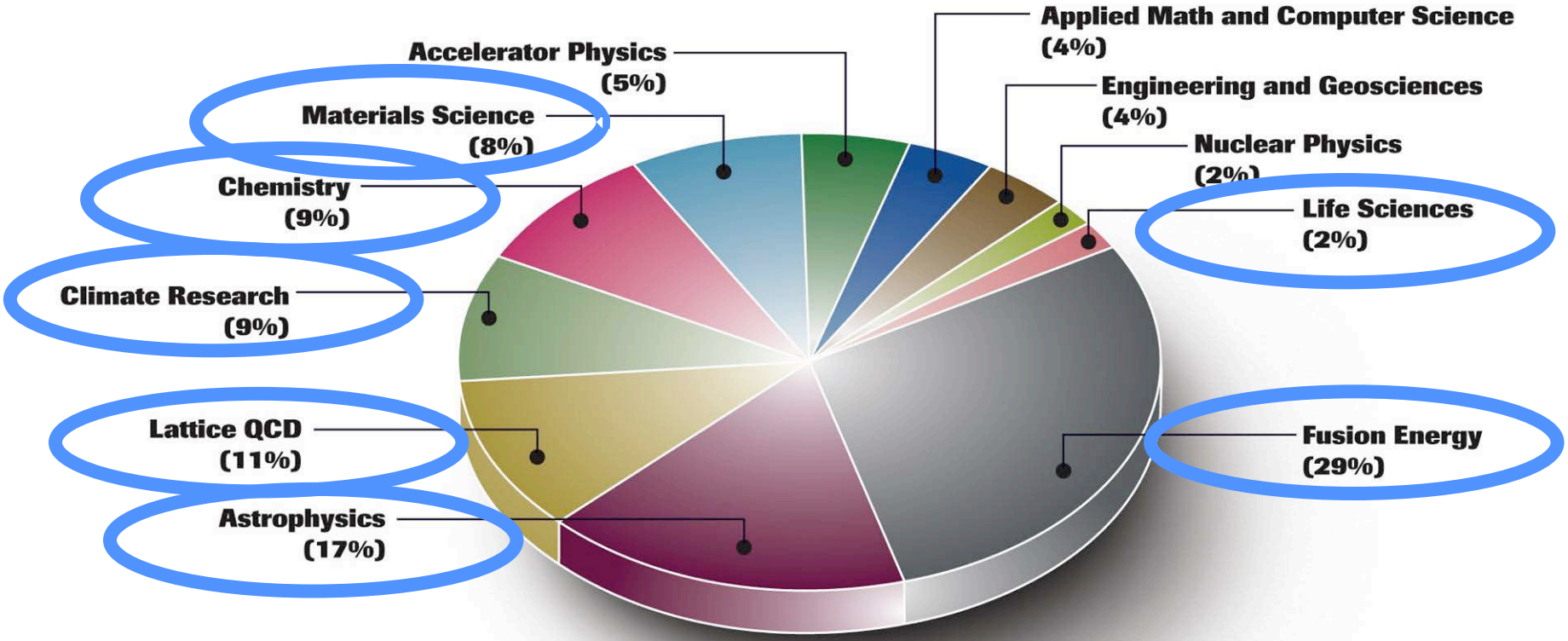
NERSC 5 Benchmarks

- **Application Benchmarks**
  - CAM3 - Climate model, NCAR
  - GAMESS - Computational chemistry, Iowa State, Ames Lab
  - GTC - Fusion, PPPL
  - MADbench - Astrophysics (CMB analysis), LBL
  - Milc - QCD, multi-site collaboration
  - Paratec - Materials science,developed LBL and UC Berkeley
  - PMEMD – Life Science, University of North Carolina-Chapel Hill
- **Micro benchmarks test specific system features**
  - Processor, Memory, Interconnect, I/O, Networking
- **Composite Benchmarks**
  - Sustained System Performance Test (SSP), Effective System Performance Test (ESP), Full Configuration Test, Throughput Test and Variability Tests

Application Summary

Application	Science Area	Basic Algorithm	Language	Library Use	Comment
CAM3	Climate (BER)	CFD, FFT	FORTRAN 90	netCDF	IPCC
GAMESS	Chemistry (BES)	DFT	FORTRAN 90	DDI, BLAS	
GTC	Fusion (FES)	Particle-in-cell	FORTRAN 90	FFT(opt)	ITER emphasis
MADbench	Astrophysics (HEP & NP)	Power Spectrum Estimation	C	Scalapack	1024 proc. 730 MB per task, 200 GB disk
MILC	QCD (NP)	Conjugate gradient	C	none	2048 proc. 540 MB per task
PARATEC	Materials (BES)	3D FFT	FORTRAN 90	Scalapack	Nanoscience emphasis
PMEMD	Life Science (BER)	Particle Mesh Ewald	FORTRAN 90	none	

Application Benchmarks represent spectrum of NERSC users



Technology Observations

- All bids were for multi core chips
  - Clock speed increasing at a much slower rate
  - The performance penalty is not as bad as we thought it might be
- Power and cooling continue to increase
  - Flop/s per $ improving faster than Flop/s per Watt or Flop/s per sf
- All proposals
  - Hybrid systems with Proprietary interconnects
  - High processor counts
  - One phase delivery - Influence of Sarbanes-Oxley?
  - Ran most to all benchmarks
  - Vertically integrated SW
- NERSC Performance predictions were accurate
  - NERSC predicted systems would be between $0.75 - $1 per peak MFlop/s and $6-8 per SSP MFlop/s.
  - Bids and selection were better than expected.

Technology Observations (cont.)

- Variety of topologies – between and within nodes
- No vector or CPU accelerated systems proposed
- Non commodity memory is very expensive
- External storage cost getting cheaper for capacity
- Delivery dates all at last moment
- All proposers can move disk drives off the single system
  - It means they all use standards compliant storage
- Declining viable bidders interested for full system and support of this size
- Is SW risk getting better? Maybe
- Efficiencies were stable and better than projected
- ESP got much better commitments
- No new technology for computer security
- No innovative technology offered

High Level NERSC 5 Features

- >15 TF Sustained System Performance
  - Geometric Mean
  - Seaborg = .89 TF
  - Bassi ~ .8 TF
- >300 TB of usable disk
- Multiple 10 GigE connections
- Many 1 GigE connections
- O(50) FibreChannel Connections
- CPUs' > 4.5 GF per core
- Multi-core sockets
- Nodes are small SMPs
- O(10,000) nodes
- ~2 GB of memory per core
- Expect programming model to remain MPI
- Linux based user environment
- Many reliability metrics
- Access to NGF via GPFS

The Phasing of NERSC-5

- **Small Test System**
  - **Summer 2006 – user access not planned**
- **Fall of 2006 - Phase 1**
  - **1/3 of compute resources**
  - **~80% of I/O infrastructure**
- **Winter 2007 – Phase 2**
  - **2/3 more compute nodes**
  - **Remaining disks and controllers**
- **Winter 2008 – option to upgrade to at least double the sustained performance**
- **Summer 2008 – Major software upgrade**
- **Winter/Spring 2009 – *option for a 1 Petaflop/s peak system – not currently in the NERSC budget***

Summary

- **NERSC continues to enable outstanding computational science through**
  - **a highly reliable, efficient, integrated production environment**
  - **provision of the whole spectrum of resources (computers, storage, networking)**
- **NERSC 5 promises to be a significant increase in production capability**